**读书报告**

61518424 王贵涛

1. **问题与解答**

我提出问题：

1. 怎么选择话题数，即k的值？

讨论结果：这个类似之前的内容，可以根据经验，也可以使用相应的指标来计算。实际应用中，我们一般根据上式，选择能使误差小于0.01（99%的信息都被保留）或0.05（95%的信息都被保留）的k值。

1. 为什么奇异值分解出来的单词-话题矩阵选择的单词表示同一个话题？

讨论结果：其实SVD的问题就在于它得出来的是计算的结果，但是不一定很好解释。它产生的话题本身也是一个抽象的表示，只是在矩阵分解的过程中这样的组合占比重很大，具有重要性，所以就作为了话题，话题的单词组合有各种各样的，最后呈现的一般是最重要的几个话题。

别人提出的问题：

1. 为什么通过奇异值分解的方法就能够直接将左奇异矩阵作为话题矩阵，后面两项作为文本在话题空间的表示？这么做有什么根据吗？

我的解答：潜在语义分析其实就是把一个矩阵分解成两个矩阵的乘积，通过奇异值分解，刚好能将矩阵分成两个。它应该只是话题的一种近似，理想中找出k个话题，但是实际上可能有很大偏差。

1. 如何直观的理解对单词-文本矩阵进行奇异值分解，可以得到单词-话题矩阵？为什么可以实现？

我的解答：单词-话题矩阵乘以话题-文本矩阵可以得到单词-文本矩阵，因此，单词-文本矩阵可以分解为单词-话题矩阵和话题-文本矩阵，通过奇异值分解的方法可以将一个矩阵分解为两个矩阵，我们将这两个矩阵分别看作单词-话题矩阵和话题-文本矩阵。

1. **下周计划安排**

看完第十七章并参加讨论。

1. **读书收获**

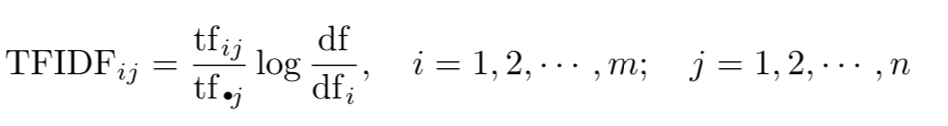
潜在语义分析（latent semantic analysis, LSA）是一种无监督学习方法，主要用于文本的话题分析。

文本信息处理中，传统的方法以单词向量表示文本的语义内容，以单词向量空间的度量表示文本之间的语义相似度。

给定一个含有n个文本的集合D，以及在所有文本中出现的m个单词的集合W。

将单词在文本中出现的数据用一个单词-文本矩阵（word-document matrix）表示，记作X。这是一个m x n矩阵，元素xij表示单词wi在文本dj内中出现的频数或权值。

权值通常用单词频率-逆文本频率（term frequency-inverse document frequency, TF-IDF）表示，其定义是：



tfij：单词wi出现在文本dj中的频数。

tf.j：是文本dj中出现的所有单词的频数之和。

dfi：含有单词wi的文本数。

df：是文本集合D的全部文本数。

单词向量空间模型的优点是模型简单，计算效率高但它也有局限性，内积相似度未必能够准确表达两个文本的语义相似度，存在一词多义性(polysemy)和多词一义性(synonymy)。

两个文本的语义相似度可以体现在两者的话题相似度上。一个文本一般含有若干个话题。如果两个文本的话题相似，那么两者的语义应该也相似。这样，基于话题的模型就可以解决上述基于单词的模型存在的问题。

在单词向量空间的文本向量xj可以通过它在话题空间中的向量yj近似表示，具体地由k个话题向量以yj为系数的线性组合近似表示X=TY。所以，单词-文本矩阵X可以近似的表示为单词-话题矩阵T与话题一文本矩阵Y的乘积形式。这就是潜在语义分析。

潜在语义对单词-文本矩阵进行奇异值分解，将其左矩阵作为话题向量空间，将其对角矩阵与右矩阵的乘积作为文本在话题向量空间的表示。